RSRE
MEMORANDUM No. 4473

ROYAL SIGNALS & RADAR ESTABLISHMENT

THE DEVELOPMENT OF THE SPEAKER INDEPENDENT *ARM* CONTINUOUS SPEECH RECOGNITION SYSTEM

Author: M J Russell

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
RSRE MALVERN,
WORCS.

RSRE MEMORANDUM No. 4473

CONDITIONS OF RELEASE

0120148 308890

****

DRIC U

****

DRIC Y

Reports quoted are not necessarily available to members of the public or to commercial organisations.

Royal Signals and Radar Establishment
Memorandum 4473

The Development of the Speaker Independent *ARM* Continuous Speech Recognition System

M J Russell
DRA (Electronics Division)
Speech Research Unit, IS2 Division
RSRE, St. Andrews, Great Malvern, England

15th January 1992

Abstract

This memorandum describes the development of a speaker independent continuous speech recognition system based on phoneme level hidden Markov models. The system is configured to recognise continuous, spoken airborne reconnaissance reports, a task which involves a vocabulary of approximately 500 words . On a test set of speech from 80 male subjects, the final system achieves a word accuracy of 74.1% with no explicit syntactic constraints.

INTENTIONALLY BLANK

Contents

INTENTIONALLY BLANK

1 Introduction

The work described in this report was conducted at the UK Speech Research Unit as part of the Airborne Reconnaissance Mission (*ARM*) continuous speech recognition project. The aim of the *ARM* project is accurate recognition of continuously spoken airborne reconnaissance reports using a speech recognition system based on phoneme-level hidden Markov models (HMMs).

Previous versions of the *ARM* system ([4, 8, 16, 7, 11, 12]) have been speaker-dependent, requiring approximately 15 to 20 minutes of speaker-specific training material (35 *ARM* reports) prior to use. Under these conditions the most recent version of the speaker-dependent *ARM* system scores an average word accuracy, without syntax, of 90.2% in laboratory tests on the 500 word *ARM* vocabulary [9]. The work reported in this memorandum is directed towards the development of a speaker-independent *ARM* system which requires no explicit speaker enrolment. Based on reported evaluations of systems in other laboratories (for example, see [13]), a performance target of 75% word accuracy, with no explicit syntactic constraints, was set for the speaker-independent system.

It was decided to concentrate initially on the development of a system for recognising speech from adult male speakers. This decision is justified by the assumption that a future speaker-independent system will involve automatic selection from multiple model sets, corresponding to different speaker types, and that the most rudimentary partition of a speaker population is likely to correspond broadly to the sex of the speaker. Therefore, in the context of this memorandum, the term "speaker-independent" should be taken to mean "male-speaker-independent".

It has been demonstrated in [13] and elsewhere that good speaker- independent performance can be achieved for tasks similar to the *ARM* task by using phoneme-level HMMs trained on task-specific speech from a large population of speakers. Hence the first stage in the development of the speaker-independent *ARM* system was the creation of a large, multi-speaker corpus. The RSRE "SI89" 321 speaker corpus, which was created for this purpose, is described in detail elsewhere [3] but that parts the corpus which is used in the current work is described in section 2. Spoken *ARM* reports from up to 61 male speakers from this corpus were used to train the most recent version of the speaker- dependent *ARM* system in order to obtain a

"baseline" speaker-independent system. A description of this baseline system, and its performance for different numbers of training speakers, is presented in section 4. The system was assessed on an "evaluation set" comprising spoken *ARM* reports from 10 male speakers, none of whom were in the training set.

The performance of the baseline speaker-independent *ARM* system as a function of number of training speakers was not entirely as predicted. It was anticipated that performance would initially increase with number of training speakers and then flatten out. For 8 of the 10 test speakers this was indeed the case, however for the remaining 2 speakers performance was extremely poor and apparently independent of number of training speakers. An investigation of the behaviour of the system for these two speakers led to the replacement of the 16 cosine coefficient front-end used in the speaker-dependent system by an 8 cosine coefficient front end. This work is described in section 7. In order to facilitate the use of different front-end parameterisations an altenative variable frame rate analysis scheme, in which VFR is applied to the filterbank representation, is introduced in section 7.2.

The next two sections report the results of routine enhancements which were made to the speaker-independent *ARM* system. In section 8 the use of fewer cosine coefficients in the acoustic front-end parameterisation suggests a re-evaluation of the use of the delta-cepstrum, which was previously considered but rejected in the speaker-dependent system ([16]). The inclusion of the delta-cepstrum in the front-end parameterisation resulted in an improvement in speaker-independent performance. Section 9 reports experiments on the use of different word transition penalties in the *ARM* recogniser. This was prompted by the observation that the errors made by the speaker-independent system were unduly biased towards word insertions. This version of the system (*SI-ARM* version 5), with variable frame rate analysis applied directly to the SRUbank representation, a delta-cepstrum-based front-end representation and an appropriately chosen word insertion penalty, scores an average word accuracy of 72.5% on the 10 speaker evaluation set.

Section 10 presents a reassessment of the two competing VFR analysis schemes in the context of *SI-ARM* version 5. Experimental results show that at this stage there is no significant difference between the performances obtained with the two competing schemes. Hence in *SI-ARM* version 5 the second scheme, in which VFR analysis is applied directly to the output of the filterbank analyser, is retained.

Section 11 summarises the evolution of the speaker- independent *ARM* system, in terms of the performance of it's various versions on the evaluation set, up to version 5 of the system. At this point it was decided to evaluate the system using a larger test set. This final test set comprises spoken *ARM* reports from 80 male speakers, none of whom were in the training or evaluation sets. The results of the final evaluation of the system, using the 80 speaker test corpus, are presented in section 12. In this final test the system scores an average word accuracy (respectively words correct) of 74.1% (respectively 84.1%) with no explicit syntactic constraints.

During the evolution of the speaker-independent *ARM* system, many experiments were conducted which are not reported as part of the mainstream development. These include experiments on the effects of details of the pronunciation dictionary on performance, and on the use of pronunciation networks which are able to accomodate alternative pronunciations of *ARM* vocabulary words. This work is not presented here.

The conclusions which have been drawn from the work are presented in section 13.

2 The "SI89" 321 Speaker Corpus

The speaker-dependent *ARM* system was developed using a corpus of 200 ARM reports spoken by each of three speakers. This corpus is clearly inadequate for the development of a speaker-independent system, hence it was necessary to record a new corpus of speech from a larger number of speakers.

The "SI89" corpus consists of recordings of speech from 321 subjects (230 male and 91 female). All of the subjects were members of RSRE staff who responded to a site notice requesting volunteers to participate in the production of the corpus. The recordings were made digitally on video cassette (44.1kHz sample rate) in a sound proof room using a Shure SM10A head-mounted microphone. Details of the recording procedure and equipment used have been presented elsewhere [6].

Each of the subjects recorded the following material:

- 3 airborne reconnaissance mission reports
- 6 "extracts" from airborne reconnaissance mission reports
- 10 sentences from a simulated air traffic control application
- 10 sentences from SCRIBE sentence set B
- 19 sequences of 4 digits

The different classes of recordings are described in [3]. Only the first types of recording are described here as it is these which are used in the current experiments.

2.1 The Airborne Reconnaissance Mission Reports

The form of the *ARM* reports has been described elsewhere [4] but is repeated here for completeness.

Texts of simulated airborne reconnaissance mission reports were created using an automatic sentence generator based on a finite state syntax and 497 word vocabulary, defined by the Royal Aerospace Establishment (RAE), Farnborough UK. A typical *ARM* report is as follows:

"Inflight report one dash alpha slash two six eight. Target map ref foxtrot kilo niner zero one two, correction two four three fife.
Sighting at zero one oh eight zulu.
New target defended strip.
Less than thirteen helicopters, type possibly hip.
Runways heading northwest wholly damaged, SAM defences to west intact.
TARWI seven eighths at two thousand, end of message"

The beginning (first three sentences) and end (final sentence) of the report specify a mission reference number, target location, time of sighting, and weather conditions respectively and are tightly structured. The remaining central part of the report, which describes what can be seen from the aircraft, is relatively free format. For the "SI89" corpus, 1000 such texts were generated. Since each subject recorded 3 reports, the total number of recorded spoken *ARM* reports in the "SI89" corpus is 963.

3 The Speaker-Independent *ARM* Pronunciation Dictionary

The vocabulary size for the *ARM* task is 497 words. These words are related to the phoneme-level symbols corresponding to the models in the model set by a speaker-independent *ARM* pronunciation dictionary. In the case of the majority of the words in the *ARM* vocabulary, the dictionary contains single baseform phonemic transcriptions. The main exceptions to this rule are the six short words "air", "at", "in", "of", "oh" and "or" which are allocated their own unique word-level symbols. This is motivated by the availability of sufficient examples of these words in the training corpus to support explicit word-level models, plus experience from the development of the speaker-dependent *ARM* system which showed that the number of insertion errors is reduced by modelling these common short words explicitely at the word level [4]. The dictionary also includes two "compound" words: "a few" and "a number" because the words "a", "few" and "number" only occur in these contexts in the *ARM* application. The individual words "a", "few" and "number" do not occur in the *ARM* dictionary.

This latter point is important. In *SI-ARM* the process which matches dictionary entries against the orthographic transcriptions of the training material searches for maximal matches. Therefore, for example, if "a" and "few" only occur in the context of "a few" and the composite "a few" is included in the dictionary, then "a" and

"few" will never get matched during training. During recognition this will result in an active word ("a"), which is represented by the single phoneme "shwa" in a context which does not occur in the training material, and is likely to lead to a large number of insertion errors.

4 The Baseline Speaker-dependent *ARM* System

The baseline speaker-independent *ARM* system was obtained by training version 7 ([4]) of the speaker-dependent *ARM* system on spoken *ARM* reports from the "SI89" corpus. *ARM* version 7 is described below for completeness.

4.1 The baseline system

4.1.1 Front-end acoustic analysis

The baseline speaker-independent system uses the $CC16$ parameterisation, which was described and evaluated in [16], with variable frame-rate analysis [12] applied after the cosine transform.

Front-end acoustic analysis in all versions of the *ARM* system is derived from the SRUbank filterbank analyser in its default configuration of 27 critical band filters spanning the range 0 to 10kHz and producing 100 frames per second. In the baseline system the feature vector $\vec{v}_t = (v_t^1, ..., v_t^{18})$ at time t is an 18 dimensional vector obtained from the SRUbank output coefficients as follows:

The mean channel amplitude $m(\vec{v}_t)$ of $\vec{v}_t$ is subtracted from each component of $\vec{v}_t$ and the resulting vector is rotated using a discrete cosine transform to obtain a new feature vector $\vec{w}_t$. The 17 dimensional vector $\vec{x}_t$ is obtained from the first 16 cosine coefficient (excluding coefficient 0) plus the average SRUbank channel amplitude. In detail:

$$x_t^d = w_t^d,\ d = 1, ..., 16$$
$$x_t^{17} = m(\vec{v}_t)$$

In the baseline system the sequence $(\vec{x}_t)$ is then subjected to a variable frame-rate analysis using the algorithm described in [12, 11] with threshold 350. This gives a new sequence $(\vec{o}_t)$. For each (variable frame-rate) time t, the 18^{th} component o_t^{18} of $\vec{o}_t$ is set equal to D_t, the number of SRUbank feature vectors which were replaced by $\vec{o}_t$ in the variable frame-rate analysis process.

4.1.2 Acoustic-Phonetic Processing

Acoustic-phonetic processing uses a set of 1495 HMMs. The model set consists of the following components:

- Four single state "non-speech" HMMs to cope with non-speech sounds in regions of the test data between spoken sentences.
- Six word-level HMMs for the commonly occuring short words "air", "at", "in", "of", "oh" and "or". The number of states in each of these word-level HMMs is equal to three times the number of phonemes in the baseform transcription of the corresponding word.
- A set of 1485 three-state triphone HMMs, one for each word-internal triphone which occurs in the *ARM* vocabulary according to the speaker-independent *ARM* dictionary.

As with earlier versions of the *ARM* system, all HMM states are identified with single multivariate Gaussian state output probability density functions sharing the same "grand" diagonal (co)variance matrix.

5 HMM Training and Recognition

5.1 Training and Test Data

The experiments described in this memorandum were conducted using training material from up to 61 male subjects from the "SI89" corpus. For each subject, only the three recordings of complete *ARM* reports were used for training. The evaluation set, which was used for testing during the development of the system, consists of three reports each from 10 male speakers who were not included in the training set. The assessment of the final system was done on a test set consisting of three reports each from 80 male speakers, none of whom were in the training or evaluation sets. The speakers in the training, evaluation and test sets are specified in table 1.

5.2 HMM Training

Monophone HMMs were obtained using training material labelled orthographically at the sentence level only. Standard sub-word HMM training procedures were used in which sentence level HMMs were constructed from phoneme-level HMMs using the dictionary of baseform transcriptions of *ARM* vocabulary words. These models were then mapped onto the sentence level acoustic data and contributions to the

Data Set	Speaker Numbers
61 Speaker Training Set	004, 005, 006, 011, 013, 014, 016, 017, 018, 020 021, 022, 023, 027, 028, 030, 033, 034, 035, 036 037, 052, 056, 058, 061, 062, 064, 065, 066, 067 072, 073, 074, 075, 076, 078, 079, 081, 084, 085 087, 089, 093, 094, 095, 097, 098, 099, 101, 102 103, 104, 105, 106, 107, 108, 109, 110, 116, 117 119
10 Speaker Evaluation Set	140, 141, 142, 144, 145, 146, 147, 148, 165, 237
80 Speaker Test Set	200, 201, 202, 204, 205, 206, 208, 209, 210, 215 216, 217, 218, 219, 220, 221, 222, 223, 225, 227 228, 230, 231, 232, 233, 234, 235, 236, 238, 239 240, 241, 242, 243, 244, 245, 246, 247, 248, 249 250, 251, 252, 253, 254, 255, 257, 258, 259, 260 261, 262, 264, 265, 266, 267, 268, 269, 273, 274 275, 276, 277, 279, 280, 281, 282, 283, 284, 285 286, 287, 288, 289, 290, 292, 293, 295, 296, 299

Table 1: *Speakers in the training, evaluation and test sets used in the experiments reported in this memorandum. The speaker numbers refer to the RSRE "SI89" Corpus*

model parameter estimates computed. For the initial iteration this mapping was linear, but for subsequent iteration the standard "forward-backward" algorithm was used.

The parameters of these context insensitive monophone HMMs were used as the initial estimates for the parameters of the set of triphone HMMs. The triphone HMMs were then optimised with respect to the training set, labelled orthographically at the sentence level, using the standard sub-word HMM training procedures. This was followed by a further three iterations of the training algorithm: the first to estimate the grand diagonal (co)variance matrix, the second to reestimate the mean vectors of the state output probability density functions given the grand (co)variance matrix, and the third to do a final reestimation of the grand (co)variance matrix. During these final three stages of training all other parameters were fixed. This "fine tuning" of the grand covariance matrix was shown to be beneficial in [7].

5.3 Recognition

Recognition was performed using a one-pass dynamic programming algorithm with beam search and partial traceback [2]. Results are presented in terms of *% words*

wrong and *% word errors*. These are computed as follows, using dynamic programming to align the true transcription of the test data with the output of the recogniser:

$$\%\ words\ wrong = \frac{S+D}{N} \times 100,$$
$$\%\ word\ errors = \frac{S+D+I}{N} \times 100$$

where N is the number of words in the test set, and S, D and I are the number of words substituted (i.e. recognised as the incorrect word), deleted and inserted respectively.

6 Performance of the "Baseline" Speaker- Independent *ARM* System

Experiments were conducted using training sets consisting of 3 complete ARM reports spoken by each of 10, 20, 30, 40, 50 and 61 male subjects. Figure 1 shows % word error with no explicit syntax as a function of number of training subjects for each of the 10 speakers in the evaluation set. It is clear from the figure that there are two modes of performance.

For the eight best speakers, recognition accuracy increases with number of training speakers for training sets with up to 40 speakers, after which it is approximately constant. The average word error for these 8 subjects with models trained on 61 speakers is 39.5%, with individual scores ranging from 58.7% to 25.8%. For the remaining two speakers (speakers 144 and 145) the performance of the system is badly degraded, with an average word error of 132.8% . Furthermore, for these speakers there is no clear correlation between number of training speakers and performance.

A detailed investigation showed that neither the acoustic data nor the annotation data for these two speakers had been corrupted, and that the poor performance was not due to errors in the recognition software. Also no obvious reason for the degraded performance is apparent from listening to the recordings: the speaking styles of these two speakers are, subjectively, no more atypical than those of the other 8 speakers.

This system, trained on 61 male speakers, will be refered to as *SI-ARM* version 1. Thus *SI-ARM* version 1 scores an average of 58.1% word error (28.3% words wrong) on the 10 male speaker evaluation set, with no explicit syntactic constraints. A breakdown of this performance against the 10 subjects in the evaluation set is included in table 2.

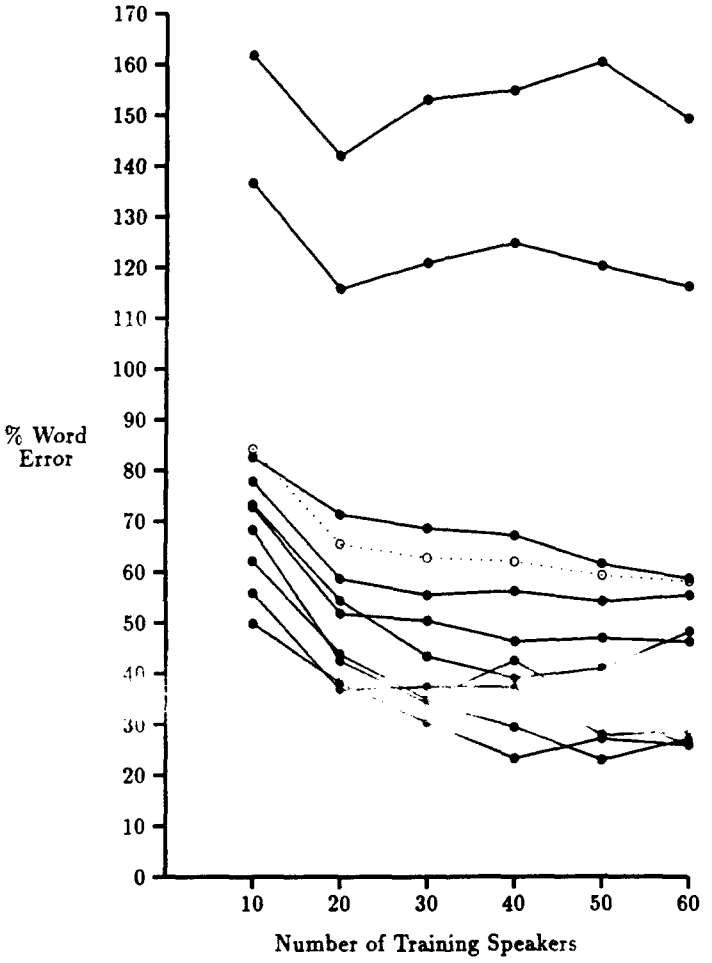

Figure 1: *Performance of the "baseline" speaker independent ARM system on the evaluation set as a function of number of training speakers (% word errors without explicit syntax). The solid and dotted lines show the scores for individual speakers and averaged over all speakers respectively.*

7 "Sheeping the Goats" : Improving Performance for Specific Subjects

As a consequence of the results presented in figure 1 an investigation was begun into the poor performance of the system for two of the test subjects. This investigation focussed on two components of the system: the variable frame rate analysis procedure and the cosine transform. The investigation of the variable frame-rate analysis procedure was motivated by the fact that the parameters of the variable frame-rate algorithm were chosen as a result of speaker-dependent experiments reported in [12, 11, 10]. The cosine transform was investigated in case some of the higher cosine coefficients should correspond to very speaker-specific properties of the speech signal.

7.1 Effect of Variable Frame-Rate Analysis

An experiment was conducted to re-assess the effect of variable frame rate analysis in the context of the speaker-independent *ARM* system. Figure 2 shows recognition accuracy for the 10 speaker evaluation set as a function of variable frame rate analysis threshold for the baseline speaker-independent *ARM* system trained on the 61 male speaker training set. The figure shows that the optimal values of the VFR threshold are similar to those for the speaker- dependent system [4]. The best performance, 57.5% word errors, is obtained with a threshold of 450, but this is not significantly better than the figure of 58.1% word errors obtained with the original VFR threshold of 350. In particular the performance is worse with the lower threshold of 250, for which fewer acoustic vectors are discarded during the VFR process.

This experiment provides strong evidence that the poor performance of the baseline speaker independent *ARM* system is not due to any inability of the variable frame rate analysis procedure to transfer successfully to the speaker-independent system.

7.2 Modifications to the Variable Frame-Rate Analysis Procedure

An important difference between the front-end processing in the "baseline" system described above and that used in the most recent version of the speaker-dependent *ARM* system is that in the latter system variable frame rate analysis is applied immediately after the SRUbank filterbank analysis and before the application of the cosine transform [10]. This has two advantages. Firstly, it improves recognition accuracy in the speaker dependent system [10]. Secondly, it allows us to assume that the metric and threshold in the variable frame-rate analysis algorithm need only be optimised for the SRUbank parametrisation, and that possible interactions

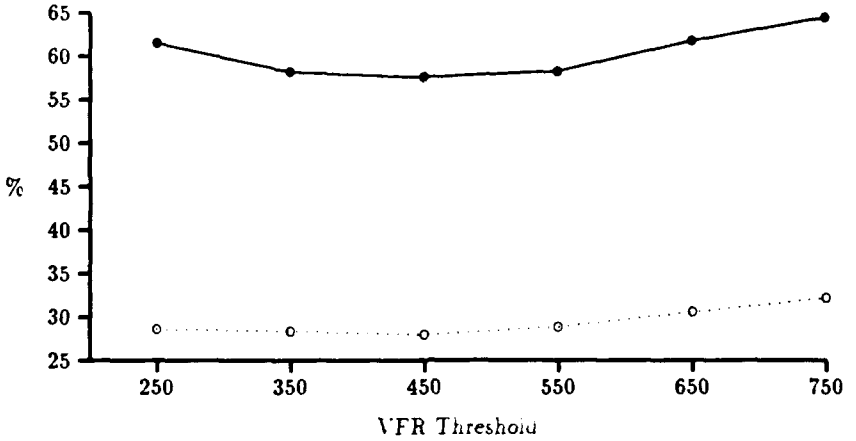

Figure 2: *Performance of the* SI-ARM *system as a function of VFR threshold for VFR analysis applied to the cosine transformed speech data (% word errors (solid line) and % words wrong (dotted line) with no explicit syntactic constraints).*

between subsequent transformation of the SRUbank representation and the variable frame-rate analysis procedure can be ignored (of course this may not be the case). For practical reasons the second advantage is particularly important in the present context, because at this point it is necessary to conduct experiments with different front-end parameterisations and we wish to avoid the overhead of re-selecting an appropriate VFR threshold for each new parameterisation. An experiment was therefore conducted to investigate the effect on performance of applying variable frame-rate analysis immediately after the SRUbank analysis. The baseline speaker independent *ARM* system. The experiment uses the same set of 61 male training speakers and 10 male test speakers as in the previous section. The results are presented in figure 3, which shows % word error and % words wrong (with no explicit syntactic constraint) for the new variable frame rate analysis scheme as a function of VFR threshold. For comparison, figure 4 shows % word error for the new and old VFR schemes as a function of the percentage of frames which remain in a typical data file after variable frame rate analysis. The new scheme, applied with an optimal threshold of 1100, results in an average word error of 61.4% and a reduction in the number of frames to 35.6% of the original. However, the best performance obtainable with the original scheme is 57.5% average word errors with 37.4% of the original data (with a VFR threshold of 450). Although the original VFR scheme gives the best performance it was decided that the new VFR scheme would be adopted during the development of the speaker-independent ARM system, because of its convenience, and that the comparison of the two VFR schemes would be repeated for the final version of the system. Hence the new VFR scheme, with a threshold of 1100, was

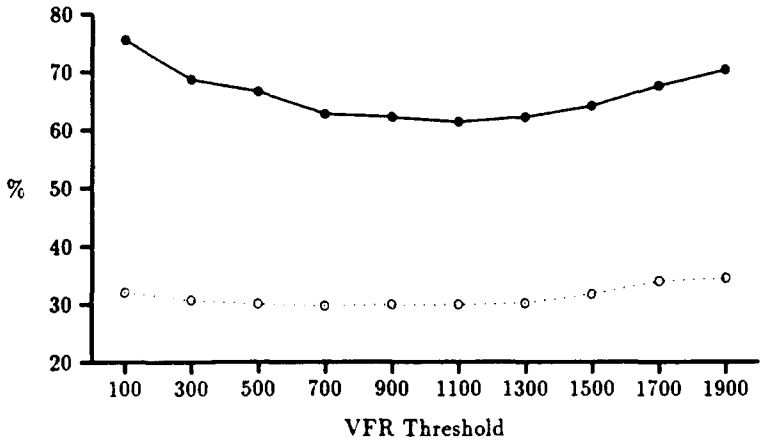

Figure 3: *Performance of the* SI-ARM *system as a function of VFR threshold for VFR analysis applied to the SRUbank data, prior to the cosine transform (% word errors (solid line) and % words wrong (dotted line) with no explicit syntactic constraints).*

adopted in the experiments described in section 7.4 and all subsequent experiments.

This version of the speaker-independent *ARM* system, in which variable frame-rate analysis is performed immediately after the filterbank analysis will be refered to as *SI-ARM* version 2.

7.3 Effect of the Cosine Transform

Figure 5 shows average values of cosine coefficients 1 to 16, after variable frame-rate analysis, computed over a single *ARM* report for each of the 10 speakers in the evaluation set. The values for speakers 144 and 145 are indicated by solid lines and those for the remaining 8 speakers by dotted lines. The figure shows that, on average, there is some separation between the values for speakers 144 and 145 and those for the remaining speakers for some of the higher cosine coefficients. Although these differences are small, they may still lead to relatively large differences in probability because the grand variances for high cosine coefficients will also be small [7]. Additional information is needed to determine whether or not these differences are significant in this sense.

In order to pursue this investigation further an experiment was conducted to explicitely measure the contributions to the observation probabilities due to individual cosine coefficients. "Forced recognition" experiments were conducted in which

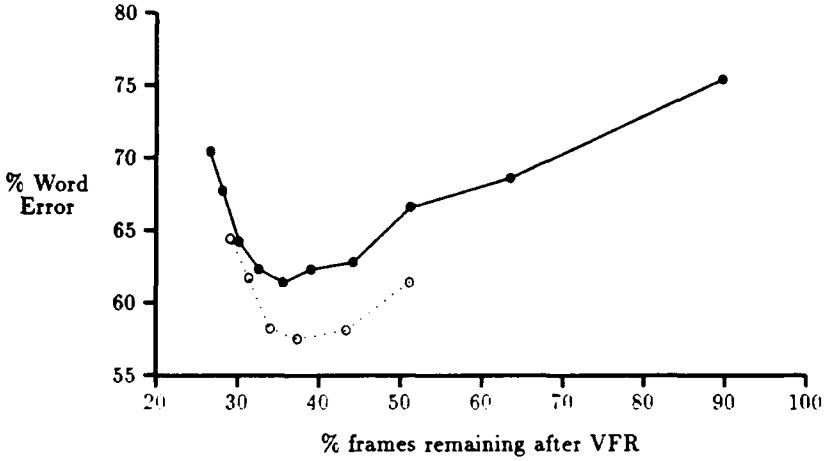

Figure 4: *Performance of the* SI-ARM *system as a function of percentage of frames remaining after VFR analysis for VFR analysis applied prior to the cosine transform (solid line) and after the cosine transform (dotted line) (% word errors with no explicit syntactic constraints).*

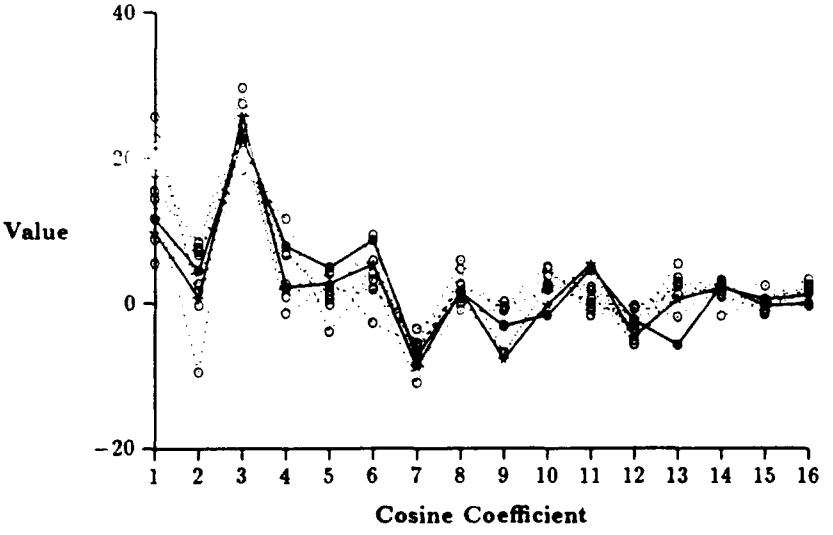

Figure 5: *Average cosine coefficient values over a single ARM report for speakers 144 (bullet, solid line), 145 (star, solid line), and the remaining speakers in the evaluation set (circles, dotted lines)*

the acoustic patterns corresponding to spoken *ARM* reports from each of the test speakers were aligned against the "correct" sequence of HMMs using the Viterbi algorithm. If $\vec{o} = \vec{o_1}, ..., \vec{o_T}$ denotes the sequence of feature vectors corresponding to a particular utterance, and $\sigma = \sigma_1, ..., \sigma_T$ is the corresponding optimal state sequence computed by the Viterbi algorithm, then the contribution $log(P_t(\vec{o}, \sigma))$ to the joint log probability of $\vec{o}$ and σ for (VFR) time t is given by:

$$log(P_t(\vec{o}, \sigma)) = -\sum_{d=1}^{18} \frac{(\vec{m_t^d} - \vec{o_t^d})^2}{(\vec{v_t^d})^2} + constant$$

$$= \sum_{d=1}^{18} log(P_t^d(\vec{o}, \sigma)) + constant$$

where $\vec{m_t}$ and $\vec{v_t}$ are the mean and variance vectors associated with state σ_t and $P_t^d(\vec{o}, \sigma)$ is the joint probability of the d^{th} cosine coefficient in $\vec{o_t}$ and state σ_t.

The contributions $P_t^d(\vec{o}, \sigma)$ due to the individual cosine coefficients are independent because of the assumption that the covariance matrix is diagonal. Figure 6 shows average values of $-log(P_t^d(\vec{o}, \sigma))$ for $d = 1, ..., 16$ computed over a single *ARM* report for each of the ten test speakers. The values for speakers 144 and 145 are joined with solid lines, while the values for the remaining eight speakers are joined with dotted lines.

The figure shows large differences between the graphs for speakers 144 and 145 and the graphs for the other speakers for cosine coefficients 9 and 13 and also some of the other higher cosine coefficients. This suggests that these coefficients are particularly sensitive to speaker-dependent factors which distinguish speakers 144 and 145 from the other test speakers. For example, the peak in the average value of $-log(P_t^9(\vec{o}, \sigma))$ for speaker 145 suggests that there is periodic structure, with period 6 channels, in the filterbank-analyser output frames for that speaker, and that this structure is characteristic of this speaker. Since the number of channels in the SRUbank representation is 27, this means that one would expect to see four equally spaced peaks in the spectrum. Observation shows that this type of structure does indeed occur in the SRUbank data for this speaker, in particular in regions of the data which correspond to the "shwa" vowel. We believe that this factor, plus the fact that this speaker displays a tendency to centralise vowels, acounts for the poor performance.

7.4 Effect of Reducing the Number of Cosine Coefficients

As a consequence of this work, the number of cosine coefficients in the acoustic front-end parameterisation was reduced from 16 to 8. Hence, since the mean channel amplitude and variable frame-rate analysis frame-count parameters were retained, the dimensionality of the resulting front-end is 10.

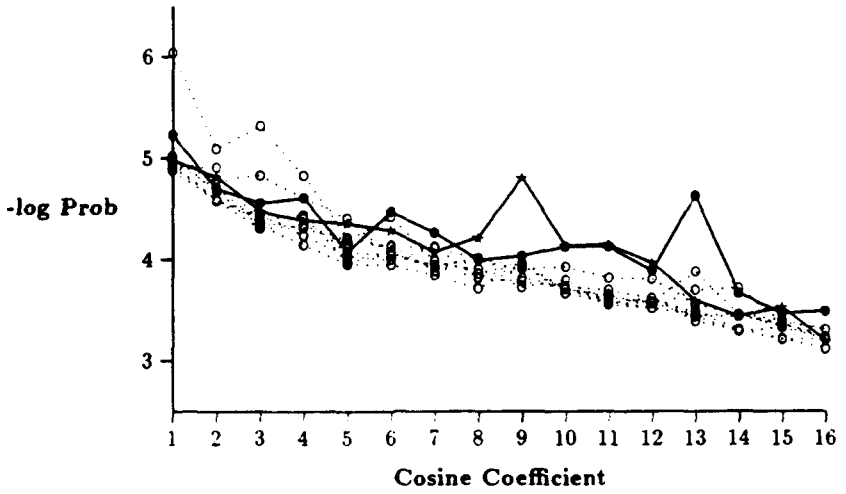

Figure 6: *Average cosine coefficient channel -log probabilities over a single ARM report for speakers 144 (bullet, solid line), 145 (star, solid line) and the remaining speakers in the evaluation set (circles, dotted lines)*

The speaker-independent *ARM* system with this modified front-end will be refered to as *SI-ARM* version 3. *SI-ARM* version 3 scores an average word error of 51.0% (25.1% words wrong) on the 10 speaker evaluation set with no syntax. A breakdown of these results against the individual speakers in the evaluation set is presented in table 2.

The effect on performance for each of the speakers in the evaluation set of moving from the original 16 cosine coefficient front-end in *SI-ARM 2* to 8 cosine coefficients in *SI-ARM 3* is shown in table 2 and figure 10 in section 11. The figure shows that the use of the lower-dimensional representation leads to a substantial improvement in the performance for speakers 144 and 145, as predicted. However, the average word error for the remaining 8 speakers increases from 40.8% with the 16 cosine coefficient front-end to 44.3%. Notice that although the performance for speakers 144 and 145 has been improved, it is still worse than that for any of the other speakers.

It is important to note that the decision to discard the top 8 cosine coefficients was made as a consequence of a study of the behaviour of the system on the evaluation set. Hence the evaluation set has been used for training and has been compromised as a true "unseen" test set. Progress during the development of the speaker independent *ARM* system will continue to be measured against this evaluation set, however the assessment of the final system will use the "unseen" test set.

8 Delta Cepstrum

In experiments with the speaker-dependent *ARM* system reported in [16] it was shown that for a front-end parameterisation based on 8 cosine coefficients, a significant improvement in performance was achieved by including time-difference, or "delta cepstrum" information. This is the $CC8\delta$ front-end described in [16]. In the speaker dependent *ARM* system the $CC8\delta$ front-end was not adopted because it was outperformed by the more simple 16 cosine coefficient front-end. However there is evidence that the use of the delta cepstrum offers significant advantages in speaker-independent recognition [13]. Hence an experiment was conducted in which a delta-cepstrum was added to the front-end described in the previous section. Using the notation from section 4, version 4 of the speaker-independent *ARM* system uses a 20 dimensional front-end parameterisation defined as follows:

$$
\begin{aligned}
o_t{}^d &= w_t{}^d,\ d = 1, \ldots, 8 \\
o_t{}^9 &= m(\vec{v_t}) \\
o_t{}^{10} &= D_t \\
o_t{}^d &= o_{t+\delta}{}^{d-10} - o_{t-\delta}{}^{d-10},\ d = 11, \ldots 20
\end{aligned}
$$

Again a recognition experiment was conducted on the 10 male subject evaluation set, using the 61 male speaker training set. The average % word errors without syntax falls from 51.0% without delta cepstrum to 36.1% with delta cepstrum. This result confirms the value of the delta cepstrum in a speaker-independent system. The results of this experiment are shown in more detail in table 2 and figure 10 in section 11.

This version of the speaker-independent *ARM* system will be refered to as *SI-ARM* version 4.

9 Word Transition Penalties

The patterns of errors in the speaker-independent *ARM* systems described above are biased towards word insertions. For example, the average word error over the 10 male speaker test set of 36.1% for version 4 of the speaker independent system can be broken down into substitution, deletion and insertion rates of 7.2%, 10.2% and 18.8% respectively. This problem is well known [5]. The standard method for balancing the word insertion and word deletion errors is to use a "word transition penalty" [15, 9]. This normally takes the form of a fixed, system-wide, "word transition probability" by which state sequence probabilities are multiplied within the Viterbi algorithm whenever a state sequence includes a transition into a new word. One can also envisage alternative schemes where the transition penalty is incurred each

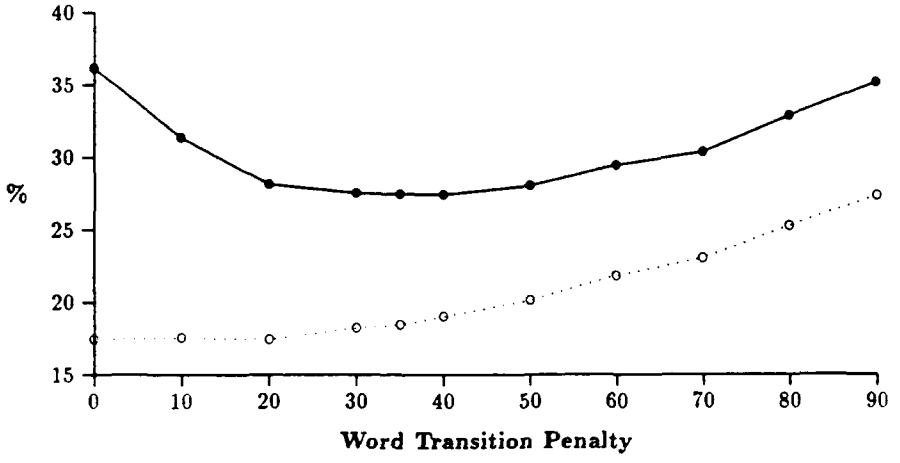

Figure 7: *Average values of % word errors (solid line) and % words wrong (dotted line) on the evaluation set for word transition penalties from 0 to 90.*

time a state sequence includes a transition into a new model, so that the "per word" penalty depends on the number of phonemes in the word. This was investigated in [9], where it was shown that these "model transition probabilities" did not perform as well as the word transition probabilities in recognition experiments with the speaker-dependent ARM system. Hence only word transition probabilities are considered in the current work. In fact, since the recognition algorithm is normally implemented in log arithmetic, it is usual to talk in terms of the word transition *penalty*, which is defined to be equal to the negative logarithm of the word transition probability.

Figure 7 shows % word errors and % words wrong as a function of the word transition penalty. It is clear from the results that the use of a word transition penalty leads to a substantial improvement in recognition accuracy. For example, with a word transition penalty of 30, the average % word error and % words wrong over the 10 speaker evaluation set is 27.5% and 18.2% respectively. This compares with 36.1% and 17.4% with a word transition penalty of 0 (i.e. with no word transition penalty). It is also evident from figure 7 that the precise value of the word insertion penalty is not critical.

Based on these results, a word transition penalty of 30 is used in all future experiments. The resulting system (*SI-ARM* version 4 with the addition of a word insertion penalty of *30*) is refered to as SI-ARM version 5.

It is interesting to note that the improvement in performance which results from the use of a word transition penalty in the current speaker- independent experiments is much greater than that observed in the speaker- dependent experiments reported in

[9].

10 Final Comparison of Alternative VFR Schemes

In section 7.2 two alternative VFR schemes were compared on an early version of the *SI-ARM* system. In the first of these VFR analysis is applied after the cosine transform and in the second it is applied after filterbank analysis but before the cosine transform. For reasons of convenience it was decided that the second scheme should be used during the development of the *SI-ARM* system, since this scheme removes the need to compute a new VFR threshold for each new representation of the acoustic signal. However, in section 7.1 the first scheme was seen to give the best performance. Therefore it was decided that the comparison of the two VFR schemes should be repeated for the final version of the *SI-ARM* system.

Figure 8 shows % word errors and % words wrong on the 10 speaker evaluation set for *SI-ARM* version 5 with VFR analysis applied to the *CC86* delta-cepstrum parameterisation. Several features of the results are of interest. First, the best word error rate obtained with this VFR scheme is 27.3%, which is not significantly different from the corresponding score (27.5%) for VFR analysis applied directly to the SRUbank representation. Second, this performance corresponds to a VFR threshold (600) which reduces the number of frames to approximately 50% of the original, which is consistent with results presented elsewhere [11]. Finally, the performance of this version of the system is much less sensitive to VFR threshold than early versions.

Based on these results it was decided to retain the scheme in which VFR analysis is applied to the SRUbank representation, prior to the application of the cosine transform.

11 Summary

Figure 9 summarises the evolution of the *SI-ARM* system in terms of recognition performance on the evaluation set. The same results are broken down against the individual speakers in figure 10 and table 2.

It is clear from figure 10 and table 2 that the performance increases for successive versions of the *SI-ARM* system are due primarily to performance improvements for particluar speakers, and that the uniform improvement across speakers is a secondary effect.

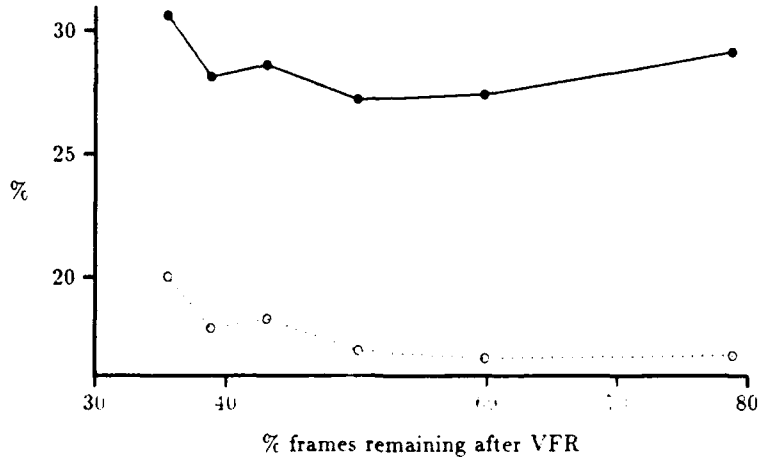

Figure 8: *Performance of the final version of the* SI-ARM *system as a function of % frames remaining after VFR analysis. with VFR analysis applied to the CC8δ representation (% word errors (solid line) and % words wrong (dotted line) with no explicit syntactic constraints).*

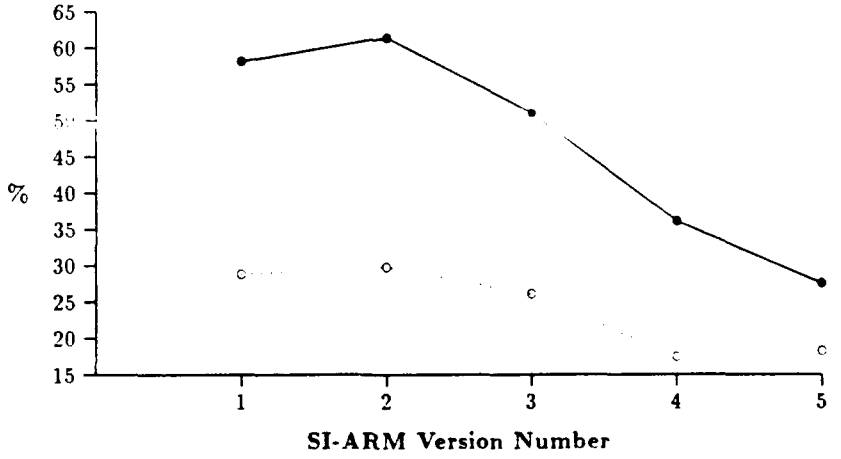

Figure 9: *Performance of all versions the* SI-ARM *system on the evaluation set (% word errors (solid line) and % words wrong (dotted line) with no explicit syntactic constraints)*

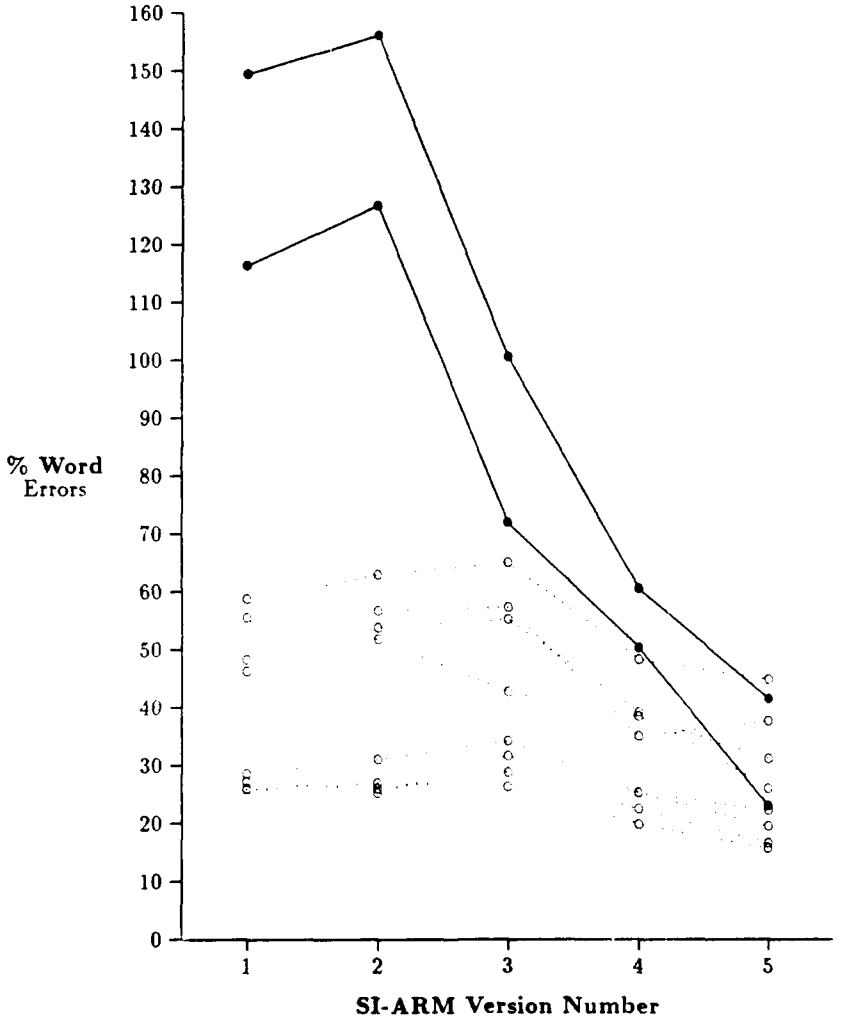

Figure 10: *Breakdown by speaker of the performance of all versions of the* SI-ARM *system on the evaluation set (speakers 144 and 145 (solid lines), and the remaining speakers (dotted lines) with no explicit syntactic constraints*

	Speaker										
	140	141	142	144	145	146	147	148	165	237	Mean
SI-ARM version 1											
%WE	25.8	48.2	27.1	149.4	116.3	25.9	55.4	28.5	46.2	58.7	58.1
%WW	16.8	20.7	12.4	66.0	51.0	9.5	32.5	15.2	27.3	32.9	28.3
SI-ARM version 2											
%WE	25.2	51.8	25.9	156.2	126.8	26.9	56.7	31.0	53.8	62.9	61.4
%WW	16.1	23.2	11.2	70.4	52.9	9.0	32.5	17.7	31.5	36.4	29.8
SI-ARM version 3											
%WE	31.6	42.7	28.8	100.6	71.9	26.3	57.3	34.2	55.2	65.0	51.0
%WW	18.7	22.6	12.4	46.9	30.7	11.4	31.8	20.9	31.5	37.1	26.1
SI-ARM version 4											
%WE	25.2	38.4	22.4	60.5	50.3	19.8	35.0	25.3	39.2	48.3	36.1
%WW	14.8	17.7	7.6	24.7	17.0	7.2	22.3	13.3	22.4	29.4	17.4
SI-ARM version 5											
%WE	19.4	31.1	16.5	41.4	22.9	15.6	37.6	22.2	25.9	44.8	27.5
%WW	15.5	21.3	8.2	23.5	11.8	8.4	30.6	24.6	19.6	30.8	18.2

Table 2: *Performance of all versions of the* SI-ARM *system on the evaluation set (%word errors (%WE) and % words wrong (%WW) with no explicit syntactic constraints)*

12 Final Evaluation of *SI-ARM*

To summarize, the final version of the Speaker-independent *ARM* system (*SI-ARM* version 5) has the following characteristics:

- **Initial front-end analysis** uses the SRUbank filterbank analyser in its default configuration of 27 critical-band spaced filters spanning frequencies up to 10 kHz and producing 100 frames per second. Each SRUbank vector is amplitude normalised, and the mean channel amplitude is stored as an additional 28^{th} channel,
- **Variable Frame-Rate analysis** is applied directly to the SRUbank output described above with a VFR threshold of 1100.
- **Secondary front-end analysis** uses a cosine transform to rotate the SRUbank data after variable frame rate analysis. The final front-end acoustic vector at (VFR) time t is a 20 dimensional delta-cepstral representation comprising:
  - cosine coefficients 1 to 8 at (VFR) time t

- mean SRUbank channel amplitude at (VFR) time t
- VFR count at (VFR) time t
- the differences between the above 10 parameters at (VFR) times $t+1$ and $t-1$

- **Acoustic-phonetic modelling** is based on a set of 1495 HMMs comprising 4 single state "non-speech" models, 6 "word-level" models of short common words and 1485 triphone models, as specified in section 4.1.2
- **Acoustic-phonetic decoding** uses the "One-pass" dynamic programming based decoding algorithm with a word insertion penalty of 30.

The final evaluation of this system uses a test set of recordings of 80 male speakers reading 3 ARM reports each. Hence the total number of *ARM* reports in the test set is 240, and the total number of words is 12,965. None of the test speakers were in the training or evaluation sets.

On this test set the above system scores 25.9% word errors (15.9% words wrong). These figures correspond to substitution, deletion and insertion rates of 4.7%, 11.2% and 10.0% respectively (see table 3).

	Percentage score	Number of words
Words correct	84.1%	10,903
Word accuracy	74.1%	
Words wrong	15.9%	2,062
Word errors	25.9%	3,355
Mismatch	4.7%	613
Deleted	11.2%	1,449
Inserted	10.0%	1,293

Table 3: *Performance of the final version the speaker-independent* ARM *system for the 80 male speaker test set (12,965 words). The system was trained on speech from 61 male training speakers. The table shows % word accuracy and % words correct with no explicit syntactic constraints*

13 Conclusions

A number of interesting conclusions can be drawn from the results presented above.

At 74.1% word accuracy, the performance of the final version of the speaker independent *ARM* system is close to the original target of 75%. This performance has been

achieved with a system which is fundamentally very simple. In particular the state output pdfs associated with the HMM states are single multivariate Gaussian pdfs with diagonal covariance matrices. Results from other laboratories suggest that this result could be improved by replacing these simple pdfs with multiple component Gaussian mixture densities.

Comparison of the final versions of the speaker-dependent and speaker-independent *ARM* systems shows that many of the empirically derived parameters, for example variable frame rate thresholds and word insertion penalties, are similar in both systems. However, a significant exception to this rule is the front-end representation. The parameterisation based on the first 16 cosine coefficients, which is used successfully in the speaker-dependent system, includes coefficients which are sensitive to speaker specific factors and hence leads to poor results for particular speakers in the speaker independent system.

A further difference between the speaker-dependent and speaker-independent systems is that the use of the delta-cepstrum, which did not result in significant improvements in recognition accuracy in the speaker-dependent system, does lead to significant improvements in the speaker-independent system.

Two alternative VFR schemes have been considered. In the first scheme VFR analysis is applied to the final (cosine transformed) representation of the speech signal, and in the second it is applied to the output of the filterbank analyser before the cosine transform is applied. Both schemes initially lead to improved performance, with the first scheme providing the best results. However, in terms of word accuracy, both the benefits of VFR and any significant differences between the two schemes diminish as the basic performance of the system increases. The results suggest that in more sophisticated systems the main (and perhaps only) benefit of VFR analysis is likely to be reduced computation.

Finally, the average word error in *SI-ARM* version 5 (27.5%) is approximately 50% of the word error achieved by the baseline system (58.1%). However, the main contribution to this improvement is a large reduction in word error for just two of the speakers in the evaluation set. For these two speakers the average word error falls from 132.8% (*SI-ARM* version 1) to 32.2% (*SI-ARM* version 5), a reduction to less that 25% of the original word error rate. Thus the improvement in performance is not uniform over all speakers in the evaluation set, but is concentrated on a relatively small subset. It would be interesting to know whether this is typical, or a feature of the particular data sets used.

References

[1] "SCRIBE - Spoken Corpus Recordings In British English : Text of Speech Material" SCRIBE Document SCRIBE-23, *Available from the Speech Research*

Unit, RSRE, Malvern.

[2] J S Bridle, M D Brown and R M Chamberlain, "A one-pass algorithm for connected word recognition", IEEE-ICASSP, 899-902, 1982.

[3] S R Browning, J McQuillan, M J Russell and M J Tomlinson, "Texts of material recorded in the SI89 speech corpus", SP4 Research Note number 142, RSRE, February 1991.

[4] M J Russell, K M Ponting, S M Peeling, S R Browning, J S Bridle and R K Moore, "The *ARM* Continuous Speech Recognition System", Proc. ICASSP'90, Albuquerque, New Mexico, April 1990.

[5] D B Paul, "A speaker-stress resistant isolated word recognizer", ICASSP'87, Dallas, TX, 1987.

[6] M J Russell, R K Moore, M J Tomlinson and J C A Deacon, "RSRE Speech Database Recordings 1983 : Part II Recordings made for Automatic Speech Recognition Assessment and Research", RSRE Report No. 84008, May 1984.

[7] M J Russell and K M Ponting, "Experiments with Grand Variance in the *ARM* Continuous Speech Recognition System", RSRE Memorandum Number 4359, 1990.

[8] M J Russell, K M Ponting, S R Browning, S Downey and P Howell, "Triphone Clustering in the *ARM* System", RSRE memorandum 4357, February 1990.

[9] K M Ponting and S M Peeling, "Word transition penalties in the *ARM* continuous speech recognition system", RSRE memorandum 4362, 1990.

[10] S M Peeling and K M Ponting, "Speaker-dependent recognition experiments using alternative front-ends with variable frame rate analysis", RSRE memorandum 4389, 1990.

[11] S M Peeling and K M Ponting, "Further experiments in variable frame rate analysis for speech recognition", RSRE memorandum 4336, 1989.

[12] K M Ponting and S M Peeling, "Experiments in variable frame rate analysis for speech recognition", RSRE memorandum 4330, 1989.

[13] K-F Lee, "Large Vocabulary Speaker-Independent Continuous Speech Recognition: the SPHINX System", PhD Thesis, Carnegie Mellon University, 1988.

[14] D B Paul, "Speaker-stress resistant continuous speech recognition", Proc ICASSP'88, New York, 1988.

[15] D B Paul, "The Lincoln Robust Continuous Speech Recogniser", Proc ICASSP'89, Glasgow, Scotland 1989.

[16] M J Russell, D Lowe, M D Bedworth and K M Ponting, "Improved Front-End Analysis in the *ARM* System: Linear Transformations of SRUbank", RSRE Memorandum Number 4358, 1990.

[17] J Wells et al., "Specification of SAM Phonetic Alphabet (SAMPA)", included in: P Winski, W J Barry and A Fourcin (Eds), "Support Available from SAM Project for other ESPRIT Speech and Language Work", The SAM Project, Department of Phonetics, University College London.

INTENTIONALLY BLANK

REPORT DOCUMENTATION PAGE

DRIC Reference Number (if known) .....

Overall security classification of sheet .....UNCLASSIFIED.....

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the field concerned must be marked to indicate the classification eg (R), (C) or (S).

Originators Reference/Report No.	Month	Year
MEMO 4473	JANUARY	1992

Originators Name and Location
RSRE, St Andrews Road
Malvern, Worcs WR14 3PS

Monitoring Agency Name and Location

Title
THE DEVELOPMENT OF THE SPEAKER INDEPENDENT
ARM CONTINUOUS SPEECH RECOGNITION SYSTEM

Report Security Classification	Title Classification (U, R, C or S)
UNCLASSIFIED	U

Foreign Language Title (in the case of translations)

Conference Details

Agency Reference	Contract Number and Period

Project Number	Other References

Authors	Pagination and Ref
RUSSELL, M J	26

Abstract

This memorandum describes the development of a speaker independent continuous speech recognition system based on phoneme level hidden Markov models. The system is configured to recognise continuously spoken airborne reconnaissance reports, a task which involves a vocabulary of approximately 500 words. On a test set of speechs from 80 male subjects, the final system achieves a word accuracy of 74.1% with no explicit syntactic constraints.

Abstract Classification (U,R,C or S)
U

Descriptors

Distribution Statement (Enter any limitations on the distribution of the document)
UNLIMITED

INTENTIONALLY BLANK